A Systematic Review of Synthetic Data Generation Techniques Using Generative AI

Year: 2024 | *Citations: 116* | *Authors: Mandeep Goyal, Q. Mahmoud*

Abstract

Synthetic data are increasingly being recognized for their potential to address serious real-world challenges in various domains. They provide innovative solutions to combat the data scarcity, privacy concerns, and algorithmic biases commonly used in machine learning applications. Synthetic data preserve all underlying patterns and behaviors of the original dataset while altering the actual content. The methods proposed in the literature to generate synthetic data vary from large language models (LLMs), which are pre-trained on gigantic datasets, to generative adversarial networks (GANs) and variational autoencoders (VAEs). This study provides a systematic review of the various techniques proposed in the literature that can be used to generate synthetic data to identify their limitations and suggest potential future research areas. The findings indicate that while these technologies generate synthetic data of specific data types, they still have some drawbacks, such as computational requirements, training stability, and privacy-preserving measures which limit their real-world usability. Addressing these issues will facilitate the broader adoption of synthetic data generation techniques across various disciplines, thereby advancing machine learning and data-driven solutions.